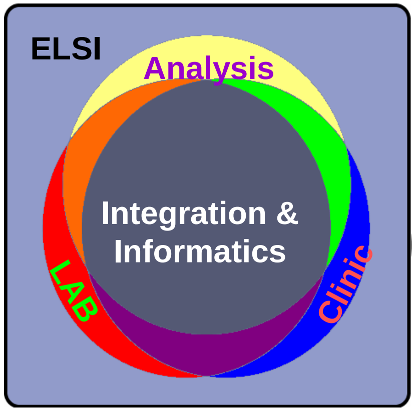
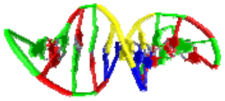
Carolina Center for Exploratory Genetic Analysis (CCEGA)



Dan Reed
reed@renci.org

Chancellor’s Eminent Professor
Vice Chancellor for IT
University of North Carolina at Chapel Hill

Director, Renaissance Computing Institute



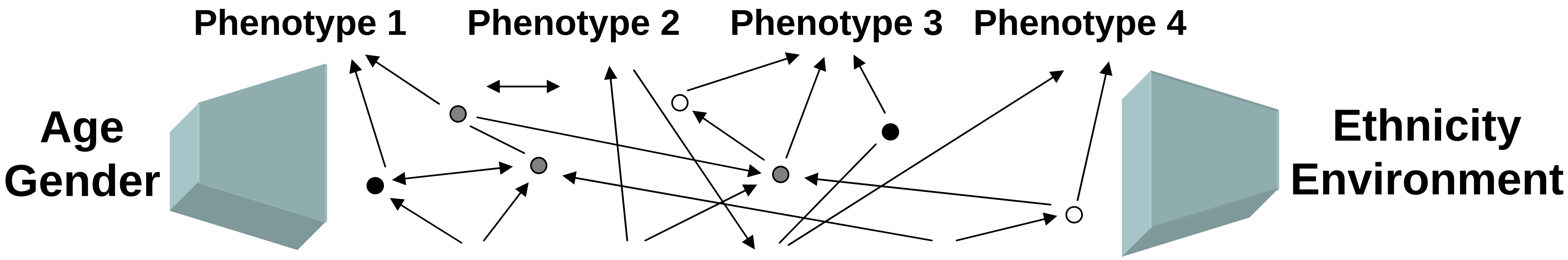
Supported in part by NIH Grant 5P20RR020751-02

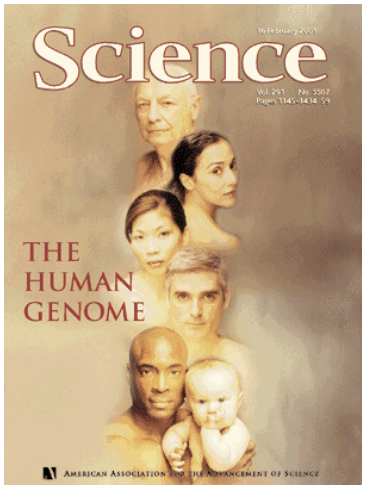
renci
renaissance computing institute

Partners and Leaders

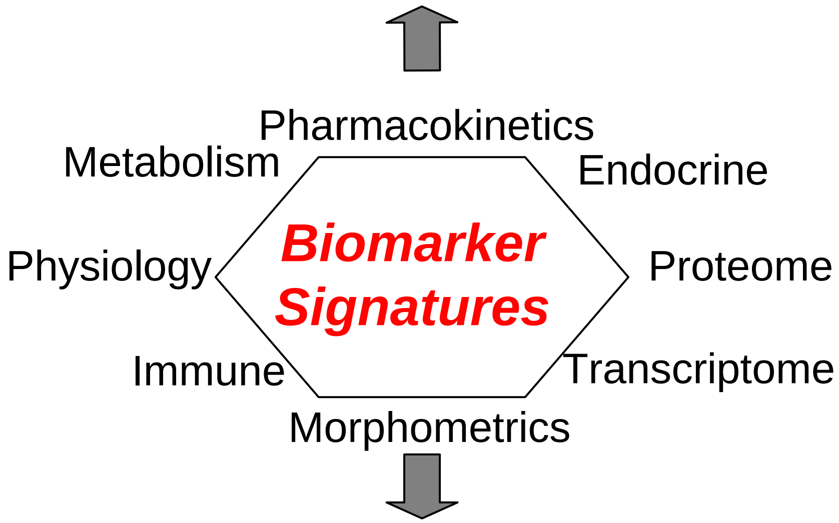
- Terry Magnuson, genetics
  - co-leader
- Kirk Wilhelmsen, genetics
  - project manager
- Jim Evans, medicine
  - ELSI
- Brad Hemminger, library and information science
  - data models and federation
- Jan Prins, computer science
  - informatics
- Fred Wright, biostatistics
- Xiaojun Guan, RENCI
  - computer science/bioinformatics
- And a host of others …

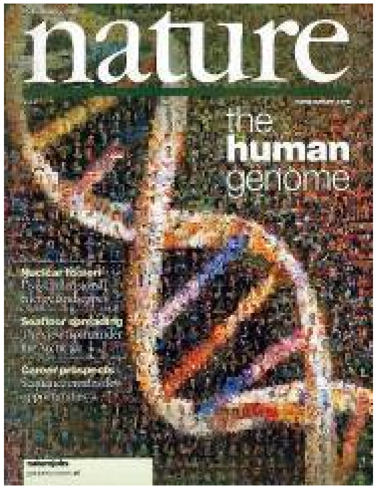
Genetics and Disease Susceptibility



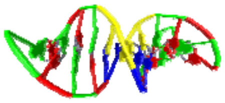


Identify Genes





Predictive Disease Susceptibility

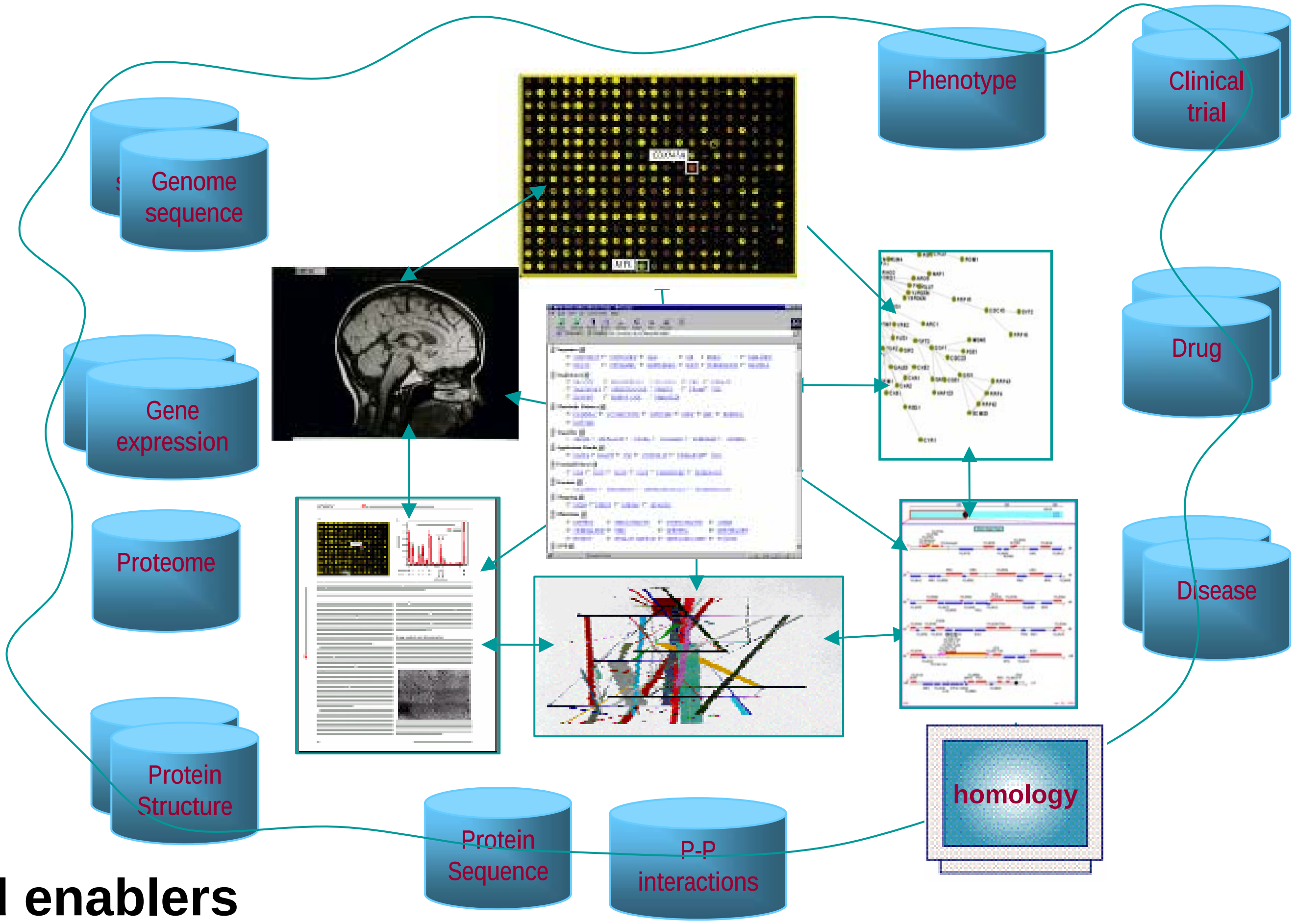


Source: David Threadgill/Terry Magnuson

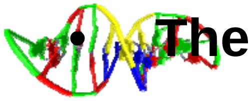
renci
renaissance computing institute

Data Heterogeneity and Complexity

Genomic, proteomic, transcriptomic, metabalomic, protein-protein interactions, regulatory bio-networks, alignments, disease, patterns and motifs, protein structure, protein classifications, specialist proteins (enzymes, receptors)



- **Many causes and enablers**
  - increased instrument resolution
  - increased storage capability
- **The challenge: *extracting insight!***



Source: Carole Goble (Manchester)

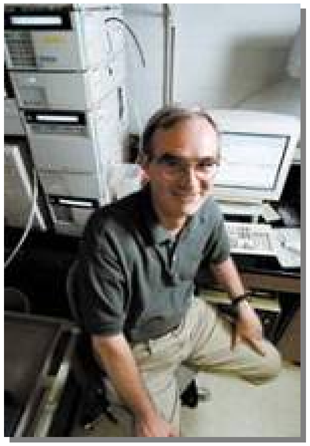
renci
renaissance computing institute

Barriers to Efficient Collaboration

- **Information Tower of Babel**
  - nomenclature and coordination
- **ELSI/IRB limitations**
  - data sharing and consent
- **Heterogeneous tools**
  - limited interoperability
  - steep learning curves
- **Culture of autonomy**
  - redundant development
    - e.g., proprietary data formats
  - best practices not always used
- **Culture gaps**
  - medicine and informatics

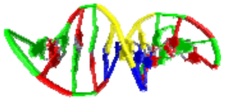
Peter Bruegel
***The Tower of Babel* (1563)**

Confluence and Opportunity



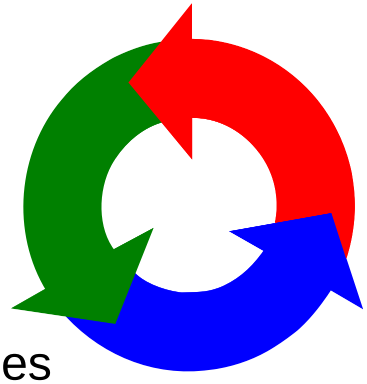
- **Center for Genome Sciences (CCGS)**
  - ten year investment of $245M
  - new center and department
  - 4 buildings and 22 faculty lines
  - advanced facilities and equipment
  - participation by multiple schools and departments
  - major gift for proteomics

- **Renaissance Computing Institute (RENCI)**
  - interdisciplinary applications of computing
  - faculty, staff and student collaborations
  - new infrastructure and capabilities
  - technology transfer and economic development
  - major state funding

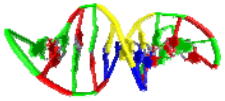


renci
renaissance computing institute

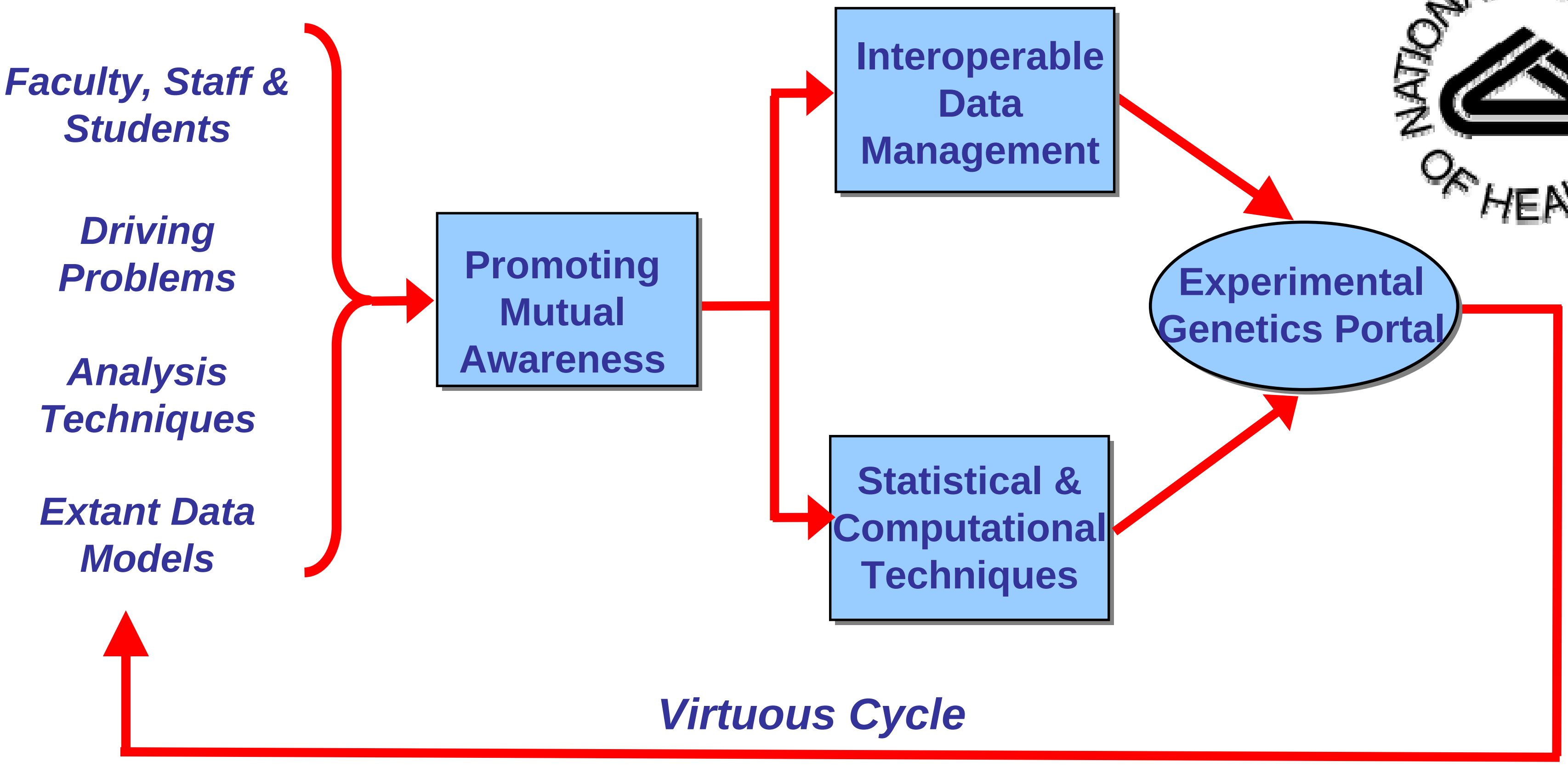
CCEGA Project Goals



- **Develop collaborative experiences and plans**
  - mutual understanding and idea generation
  - shared needs and activities
- **Deliverables and activities**
  - develop a protocol for prospective studies
    - using ongoing studies as examples to define best practices
    - Carolina Cohort
  - develop a prototype informatics infrastructure
    - data models, methods, tools and portals
  - demonstrate the utility of data mining
    - applied to established project(s)
  - facilitate use of best practices for existing projects
  - develop an environment for cross training and education
    - formal and informal education touching project participants and trainees
- **Catalyze new genetics research**



Carolina Center for Exploratory Genetic Analysis (CCEGA)



Faculty, Staff & Students

Driving Problems

Analysis Techniques

Extant Data Models

Promoting Mutual Awareness

Interoperable Data Management

Statistical & Computational Techniques

Experimental Genetics Portal

Virtuous Cycle

Interdisciplinary Research & Education

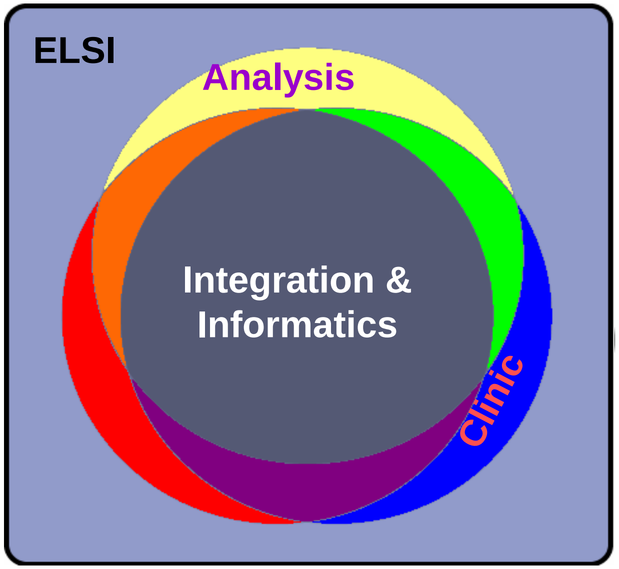
CCEGA Participation Snapshot

- **Coordination team**
  - Terry Magnuson, CCGS
  - Kirk Wilhelmsen, CCGS
  - Dan Reed, RENCI
  - Alan Blatecky, RENCI
- **Eleven departments/institutes**
  - Biostatistics
  - Cancer Center
  - Genetics
  - Computer Science
  - Epidemiology
  - Genetics
  - Health Science Library
  - Information and Library Science
  - Pharmacy
  - RENCI
  - Statistics
- **Campus wide support**
  - from many sources

- **Example project participants**
  - Brad Hemminger, Information & Library Science
  - James Evans, Genetics
  - Kevin Gamiel, RENCI
  - Xiaojun Guan, RENCI
  - Barrie Hays, Health Science Library
  - Clark Jefferies, RENCI
  - Ethan Lange, Genetics
  - Andrew Nobel, Statistics
  - Karen Mohlke, Genetics
  - Kari North, Epidemiology
  - Susan Paulsen, Computer Science
  - Fernando Manuel Pardo, Genetics
  - Charles Perou, Cancer Center
  - Lavanya Ramakrishnan, RENCI
  - Jan Prins, Computer Science
  - Patrick Sullivan, Genetics
  - Lisa Susswein, Cancer Center
  - David Threadgill, Genetics
  - Alexander Tropsha, Pharmacy
  - K.T.L. Vaughan, Health Science Library
  - Fred Wright, Biostatistics
  - Wei Wang, Computer Science
  - Fei Zou, Biostatistics

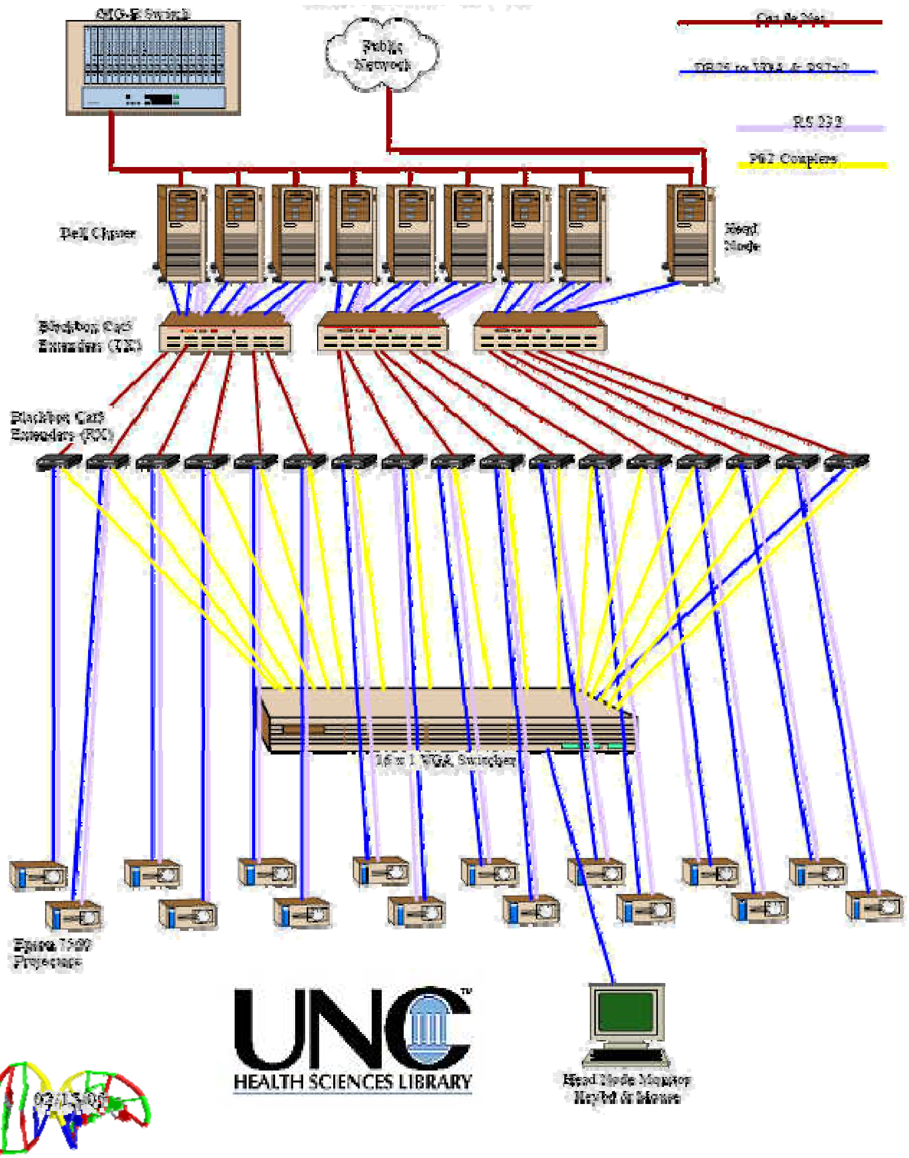
Formal CCEGA Activities

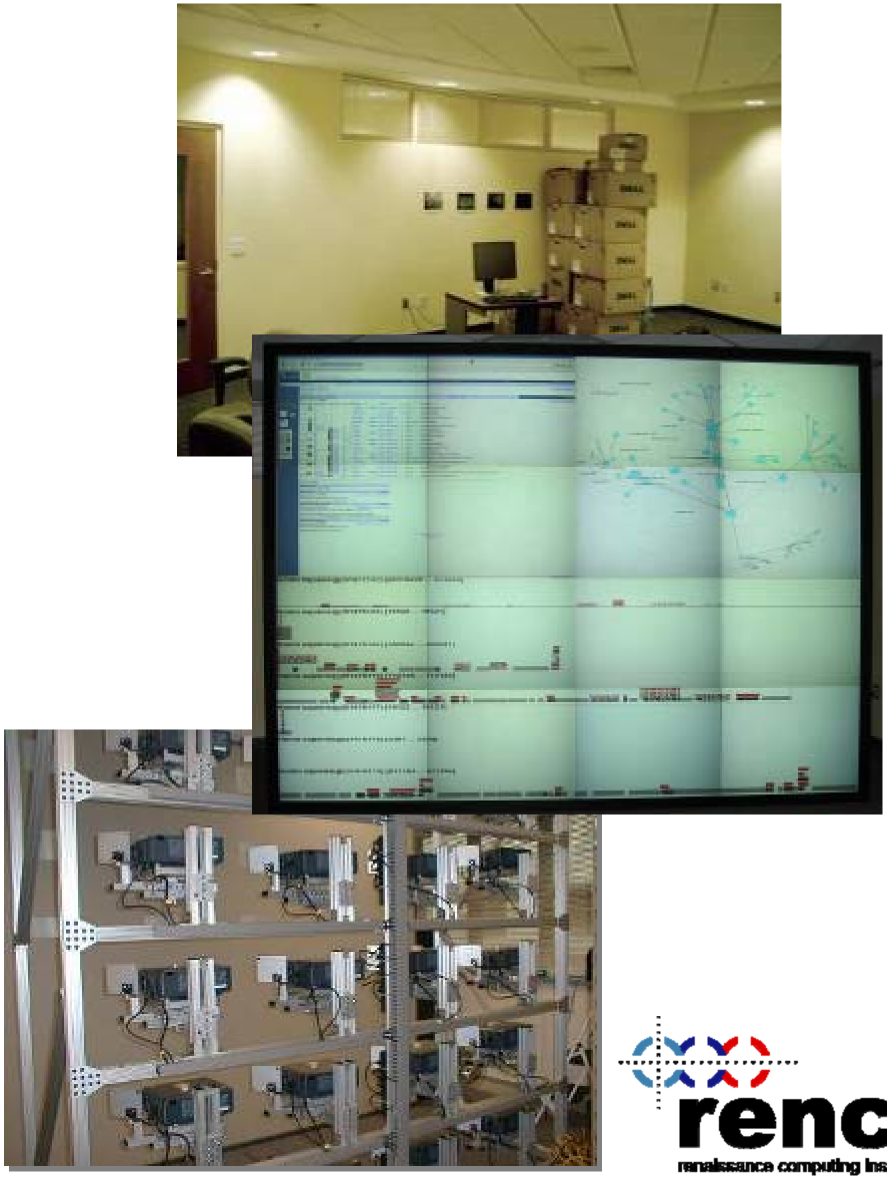
- **Workshops**
  - genetics and disease
  - analysis methods
- **Cross-disciplinary tutorials**
  - genotyping
  - XML and data representations
- **Three major working groups**
  - ELSI, analysis and informatics
- **Software prototyping**
  - portals and data model planning
- **Management group**
  - planning and strategy



www.renci.org/research/ccega

Health Science Collaboration Facility





UNC HEALTH SCIENCES LIBRARY

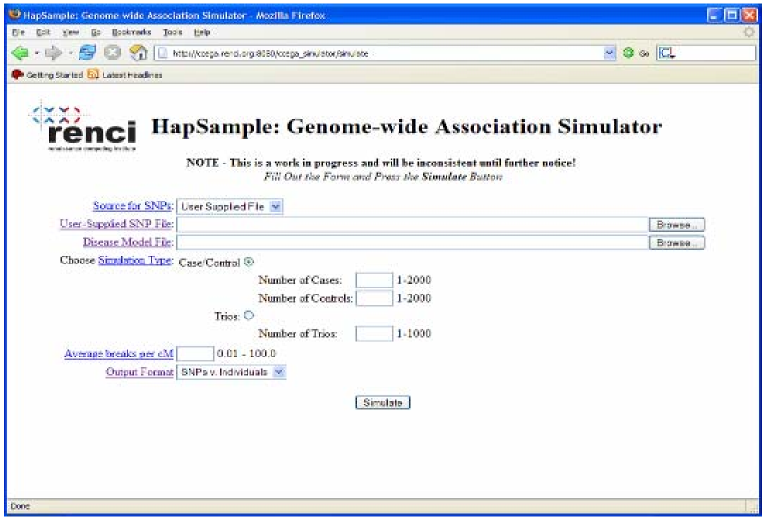
03/15/05

renci
renaissance computing institute

CCEGA HapMap Simulator

- **Resample from HapMap haplotypes**
  - create individuals with statistical properties of data
  - recombine and adjust
    - biased SNP selection and sample size
- **Model disease**
  - create large populations with families/select individuals
  - disease model can be complex
    - involving multiple loci
- **Enable analysis bakeoff**
  - five data sets simulated with 500K SNPs
    - trait caused by common sequence variants
    - each data set has 5000 cases/5000 controls
      - common versus rare traits
      - independent versus additive versus epistatic
      - variation in effect size and allele frequency
  - blind analysis by five UNC groups
    - computer science, applied math
    - biostatistics, pharmacy and genetics

nature
OPTOELECTRONICS
Germanium boost for silicon chips
LAW OF THE JUNGLE
Don't ask a chimpanzee for help
MEN OF LETTERS
If Darwin and Einstein had e-mail...
THE HAPMAP PROJECT
Chapter and verse on human genetic variation
NATUREJOBS
Biodefence boom


HapSample: Genome wide Association Simulator - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
Go
Getting Started
Latest Headlines
renci
Done

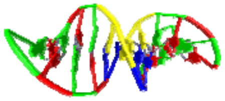
http://ccega.renci.org:8080/ccega_simulator/simulate

HapSample: Genome-wide Association Simulator

NOTE - This is a work in progress and will be inconsistent until further notice!
*Fill Out the Form and Press the **Simulate** Button*

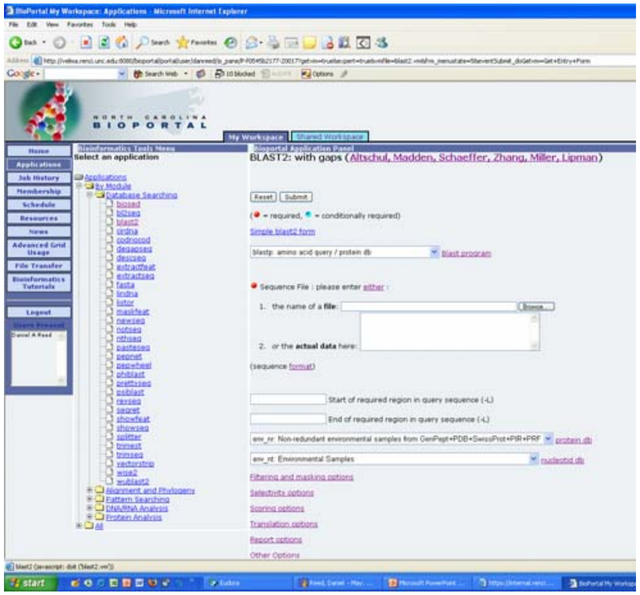
Source for SNPs: User Supplied File
User-Supplied SNP File: Browse...
Disease Model File: Browse...
Choose Simulation Type: Case/Control ●
Number of Cases: 1-2000
Number of Controls: 1-2000
Trios: ○
Number of Trios: 1-1000
Average breaks per cM 0.01 - 100.0
Output Format SNPs v. Individuals

Simulate



Carolina/CCEGA Bioportal

- **Three overlapping target groups**
  - undergraduate education
  - graduate education and research
  - academic/industrial research
- **Features**
  - access to common bioinformatics tools
  - extensible toolkit and infrastructure
    - OGCE and National Middleware Initiative (NMI)
    - leverages emerging international standards
  - remotely accessible or locally deployable
  - packaged and distributed with documentation
- **National reach and community**
  - NSF TeraGrid deployment
    - science gateway
- **Education and training**
  - hands-on workshops
    - clusters, Grids, portals and bioinformatics



Select an application

- Applications
  - By Module
    - Database Searching
      - biosed
      - bl2seq
      - blast2
      - cirdna
      - codnocod
      - degapseq
      - descseq
      - extractfeat
      - extractseq
      - fasta
      - lindna
      - listor
      - maskfeat
      - newseq
      - notseq
      - nthseq
      - pasteseq
      - pepnet
      - pepwheel
      - phiblast
      - prettyseq
      - psiblast
      - revseq
      - seqret
      - showfeat
      - showseq
      - splitter
      - trimest
      - trimseq
      - vectorstrip
      - wise2
      - wublast2
    - Alignment and Phylogeny
    - Pattern Searching
    - DNA/RNA Analysis
    - Protein Analysis
  - All

BLAST2: with gaps (Altschul, Madden, Schaeffer, Zhang, Miller, Lipman)

Reset Submit

(● = required, ● = conditionally required)

Simple blast2 form

blastp: amino acid query / protein db — blast program

● Sequence File : please enter either :

1. the name of a **file**: Browse...
2. or the **actual data** here:

(sequence format)

Start of required region in query sequence (-L)

End of required region in query sequence (-L)

env_nr: Non-redundant environmental samples from GenPept+PDB+SwissProt+PIR+PRF — protein db

env_nt: Environmental Samples — nucleotid db

Filtering and masking options

Selectivity options

Scoring options

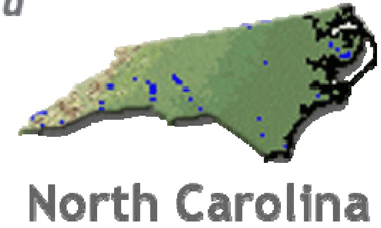
Translation options

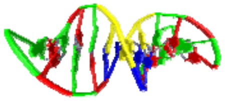
Report options

Other Options

blast2 (javascript: doit ('blast2.vm'))

TeraGrid

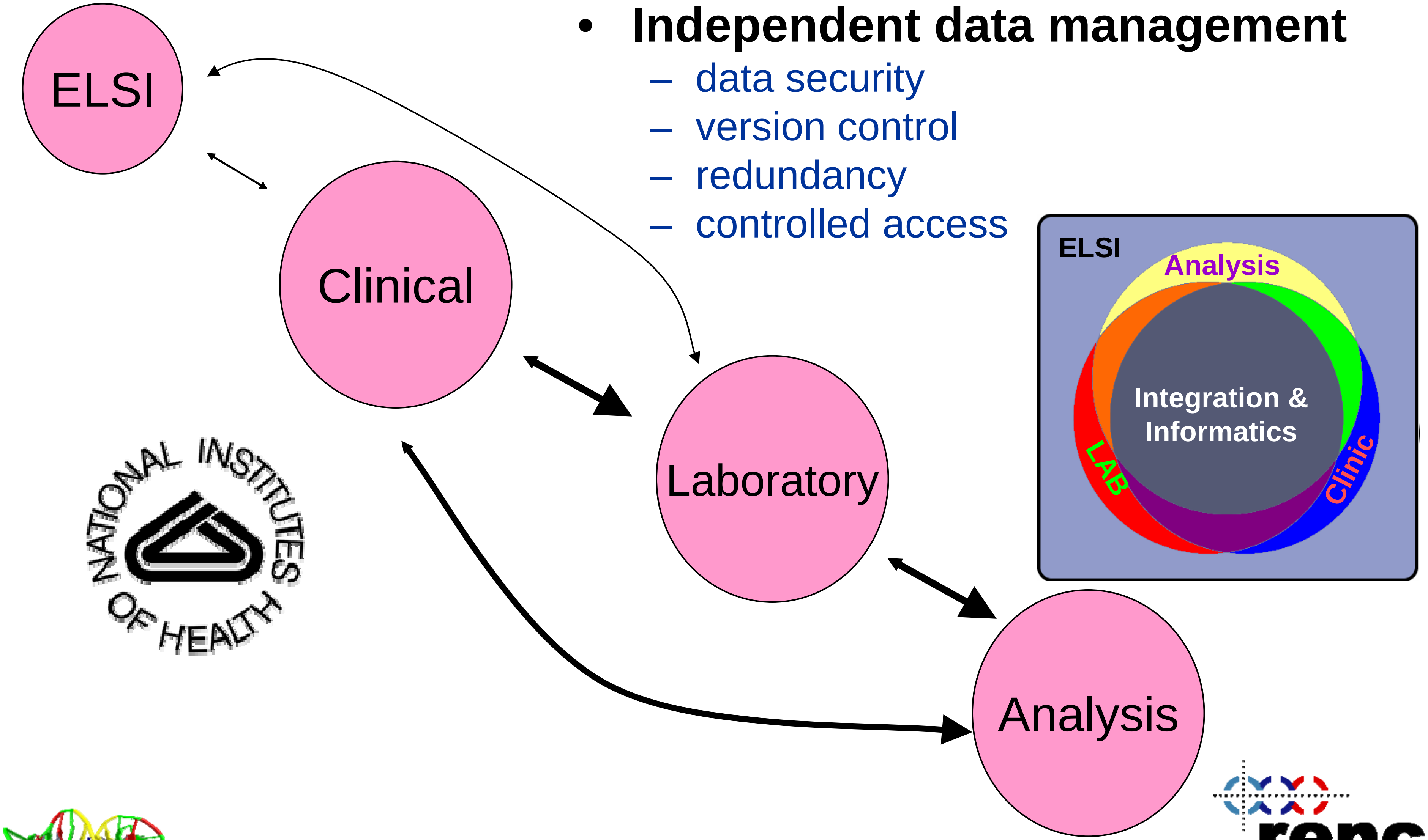




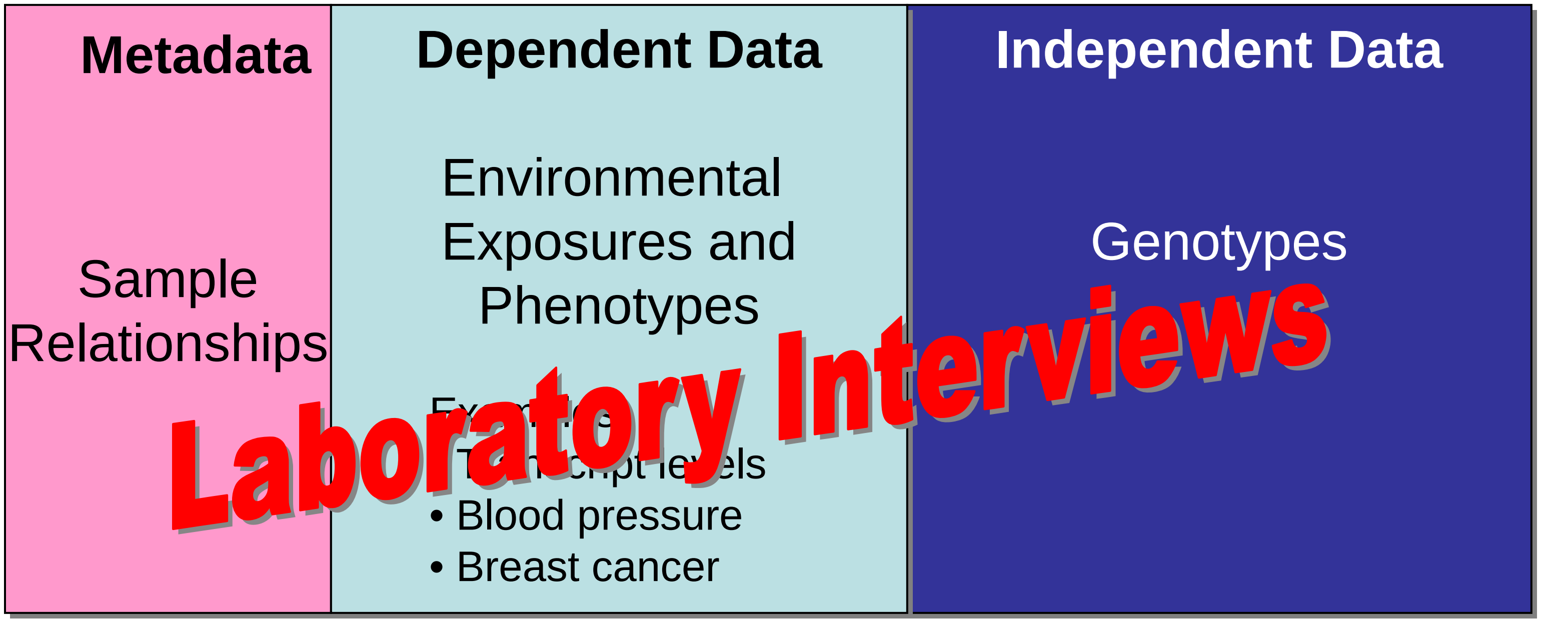
renci
renaissance computing institute

Data: From Lab and Clinic to Analysis

- **Independent data management**
  - data security
  - version control
  - redundancy
  - controlled access



Genetic Data: Conceptually a Matrix

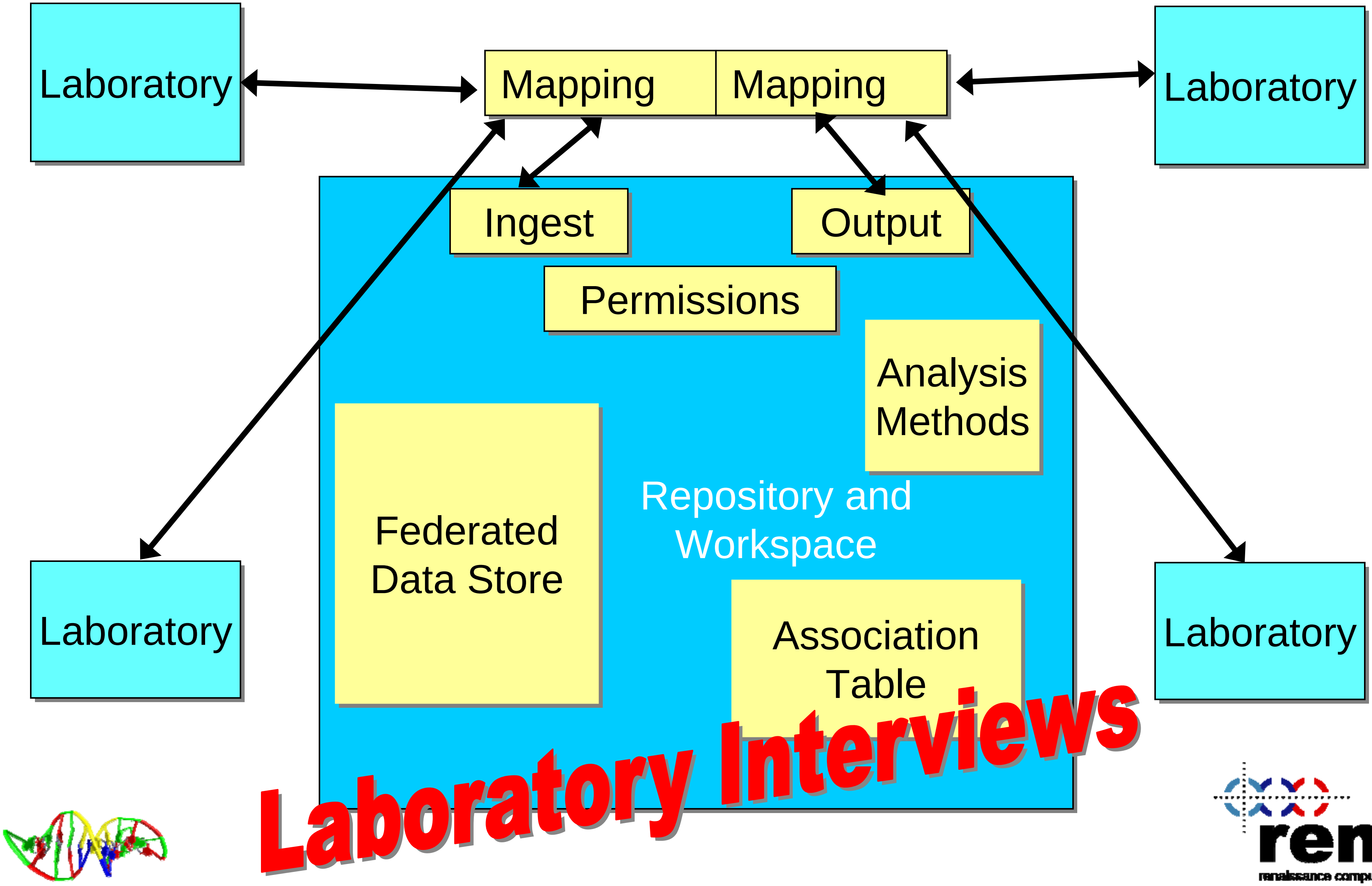


Metadata	Dependent Data	Independent Data
Sample Relationships	Environmental Exposures and Phenotypes Examples: • Transcript levels • Blood pressure • Breast cancer	Genotypes

Laboratory Interviews

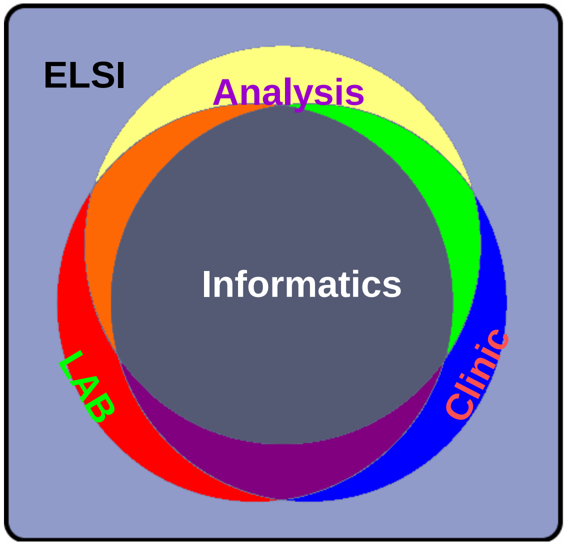
- **Rows: data on individuals**
- **Columns: multiple data values on an individual**

Federated CCEGA Data Model



ELSI Integration

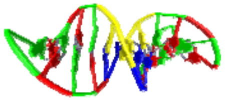
- **Novel ELSI issues from exploratory analysis**
  - practical research needs and subject rights
  - unanticipated results of exploratory analyses
  - possible unforeseen clinical implications
  - Investigator “ownership” issues
- **Outcomes**
  - overarching IRB designed
    - ensure ability to pursue such studies
  - education and engagement



ELSI considerations must be integrated throughout the entire process from study design to data/sample collection, storage, analysis and disclosure

Our Long Term Vision of Success

- **National community representation**
  - driving genetics problems and experiences
  - infrastructure testing and validation
- **Multidisciplinary collaboration**
  - biomedical and informatics researchers
  - software developers
- **National infrastructure and communities**
  - distributed and federated
    - customizable to local needs
  - interoperable and shared
- **The “Virtual Observatory” astronomy model**
  - standard tools
  - metadata and data models
  - virtual community



NVO

renci
renaissance computing institute